AD-A049 306 MINNESOTA UNIV MINNEAPOLIS DEPT OF THEORETICAL STATISTICS F/G 12/1
NEAREST NEIGHBOR RULES FOR STATISTICAL CLASSIFICATION BASED ON --ETC(U)
MAY 77 S D GUPTA, H E LIN DAAG29-76-G-0038

UNCLASSIFIED UMN/DTS/TR-285 ARO-13149.5-M NL

1 of 1
ADA049 306

END
DATE
FILMED
3-78
DDC

ARO 13149.5-M

12
B.S.

UNIVERSITY OF MINNESOTA

SCHOOL OF
STATISTICS

6

NEAREST NEIGHBOR RULES FOR STATISTICAL CLASSIFICATION BASED ON RANKS.

by

10 Somesh Das/Gupta*
University of Minnesota

and

Hsien Elsa/Lin*
National Cheng-Chi University, Taiwan

9 Technical Report, No. 285

18 ARO

14 UMN/DTS/TR-285

19 13149.5-M

11 5 May 77

12 17p.

*This research was supported by a grant from the Mathematics Division, U.S. Army Research Office, Durham N.C. Grant no. DAAG 29-76-G-0038

15

Key Words. Classification, rank nearest neighbor rule, multi-stage rule, asymptotic probability of misclassification, estimation of error probabilities.

AMS classification 62G10, 62H30

410 022

LB

1. Introduction.

The nearest neighbor (NN) rule for classifying an observation Z into one of two given populations (or, classes) π_1 and π_2 was first introduced by Fix and Hodges [3]. The rule may be described as follows. Let $(X_1,\ldots,X_{n_1})$ and $(Y_1,\ldots,Y_{n_2})$ be random (training) samples from π_1 and π_2, respectively. Using a distance function d rank the distances of all the observations from Z. Classify Z into the population to which the nearest neighbor of Z belongs. This rule was also studied by Cover and Hart [2] based on an identified training sample from a mixture of π_1 and π_2.

We shall first suggest a rule which uses the above idea in terms of the ranks of the observations in the pooled sample (including Z). The rule is specially useful when the observations are indeed available only in terms of their ranks. The rule described below will be termed as the "rank nearest neighbor" (RNN) rule.

Pool the observations X_i's, Y_j's and Z and note their ranks. (i) If Z is either the smallest or the largest observation classify Z into the class of its nearest neighbor. (ii) If both the left-hand and the right-hand neighbors (denoted, respectively, by U_1 and V_1) of Z belong to the same class, classify Z into that class. (iii) If U_1 and V_1 belong to different classes, classify Z into either of the two classes with probabilities $\frac{1}{2}$ and $\frac{1}{2}$. (We shall call this a "tie".)

In Section 2 the asymptotic (as $n_1,n_2\to\infty$) values of the probabilities of misclassification (PMC) of the RNN rule are derived. It turns out that these asymptotic values are the same as the corresponding

asymptotic PMC's of the NN rule (see [3]).

To reduce the chance of randomization in the RNN rule we consider a multi-stage version as follows. If the first-stage RNN rule (described above) leads to a tie we delete the two observations corresponding to U_1 and V_1, and apply the first-stage rule to the remaining observations. We proceed this way and move to the next stage whenever a tie occurs, and apply the first-stage rule deleting all the observations that correspond to the left-hand and the right-hand neighbors in the previous stages. The M-stage RNN rule is defined to be the one which terminates at the Mth stage (and allows for a tie in this final stage). In Section 3 the asymptotic PMC's of the M-stage RNN rule are derived.

The above rule can also be described in terms of tolerance regions based on the pooled training sample. The basic idea was suggested by Anderson [1].

We shall denote the c.d.f's of X_i and Y_j by F_1 and F_2, respectively and we shall assume that F_i possesses a density function f_i with respect to the Lebesgue measure. It is also assumed that the density of Z is either f_1 or f_2.

2. Asymptotic PMC's of the one-stage RNN rule.

The following lemma leads us to assume, without loss of generality at least for asymptotic results, that the right-hand and the left-hand neighbor of Z at the M-th stage (denoted by U_M and V_M, respectively) are well-defined. Let $n = \min(n_1, n_2)$.

Lemma 2.1 If $M/n \to 0$ as $n \to \infty$, the probability (under either $Z \sim f_1$ or $Z \sim f_2$) that there are at least M observations to the right of Z and at least M observations to the left of Z in the training sample for all

sufficiently large n is one.

Proof. Since F_i's are continuous, the probability that either $0 < F_1(Z) < 1$ or $0 < F_2(Z) < 1$ occurs is one. Suppose, in particular, $0 < F_1(Z) < 1$. It is then sufficient to prove that the probability of the event stated in the lemma conditioned by $Z = z$ is one for all z such that $0 < F_1(z) < 1$. Define

(2.1) $$W_i = I_{(z,\infty)}(X_i),$$

where I is the indicator function. Then $E(W_i) = 1 - F_1(z) > 0$. By the strong law of large numbers, we have

(2.2) $$P[\sum_{i=1}^{n_1} W_i/n_1 \to E(W_1) \text{ as } n_1 \to \infty] = 1.$$

Since $M/n_1 \to 0$,

(2.3) $$P[\sum_{i=1}^{n_1} W_i \geq M \text{ for all sufficiently large } n] = 1.$$

The corresponding result for the left-hand neighbor of Z can be proved similarly.

Next we shall prove that U_M and V_M tend to Z almost sure as $n \to \infty$.

Lemma 2.2. Given that Z is distributed as F_1, both U_M and V_M converge to Z almost sure as $n_1 \to \infty$ and $M/n_1 \to 0$.

Proof. Let

$$S_1 = \{z\colon\ F_1(z+\varepsilon) - F_1(z) > 0,\ F_1(z) - F_1(z-\varepsilon) > 0 \text{ for all } \varepsilon > 0\}$$

Then

$$P(Z \in S_1) = 1.$$

This follows from the fact that the set of intervals in which F_1 is

constant is at most countable. Thus the set of endpoints of these intervals has F_1-measure zero, since F_1 is continuous. Thus for $z \in S_1$

(2.4) $F_1(z) - F_1(z-\epsilon) > 0$

for every $\epsilon > 0$. We shall now prove that given $Z = z \in S_1$, $U_M \to Z$ a.s.

(2.5) $P[U_M < z - \epsilon] \leq P[W < M]$,

where W is the number of X_i's in $(z-\epsilon, z)$. Given $\epsilon > 0$, n_1 can be chosen sufficiently large so that

(2.6) $[F_1(z) - F_1(z-\epsilon)] - M/n_1 > \eta > 0.$

Hence

(2.7) $P[W < M] < \exp(-2n_1\eta^2)$

for all sufficiently large n_1. Hence $U_M \to z$ a.s. Similarly, it can be shown that $V_M \to z$ a.s. as $n_1 \to \infty$ for $z \in S$. The lemma now follows easily.

Let U_i and V_i be the left-hand and the right hand neighbors of Z at the ith stage. Define

$$\varphi_i \equiv \varphi_i(Z; X_j\text{'s}, Y_\ell\text{'s}; j = 1,\ldots,n_1; \ell = 1,\ldots,n_2).$$

$$(2.8) \quad = \begin{cases} 1, & \text{if both } U_i \text{ and } V_i \text{ are X-observations, or Z is an extreme observation at the ith stage and its NN is an X-observation,} \\ \frac{1}{2}, & \text{if } U_i \text{ and } V_i \text{ belong to different classes} \\ 0, & \text{otherwise.} \end{cases}$$

Let A_i be the event that both U_i and V_i are well-defined at the ith stage.

The conditional probability of deciding that Z comes from π_1 using the one-stage RNN rule, given $Z = z$, is

$$\pi^{(1)}(z; n_1, n_2) \equiv E[\varphi_1 | Z=z]$$

$$(2.9) \quad = E[\varphi_1 I_{A_1^c} | Z=z] + E[\varphi_1 I_{A_1} | Z=z]$$

However

$$(2.10)\qquad E[\varphi_1 I_{A_1^c} | Z=z] \le P(A_1^c \mid Z=z) \to 0$$

by Lemma 2.1 for almost all z. Now note that

$$\mathcal{E}[\varphi_1 I_{A_1} | Z=z]$$

$$= P[(\varphi_1 = 1) \cap A_1 | Z=z] + \tfrac{1}{2}P[(\varphi_1 = \tfrac{1}{2}) \cap A_1 | Z=z].$$

$$(2.11)\qquad = E\, P^{(11)}_{n_1,n_2}(U_1,V_1,z) + \tfrac{1}{2} E\, P^{(10)}_{n_1,n_2}(U_1,V_1,z),$$

where

$$(2.12)\qquad P^{(11)}_{n_1,n_2}(u,v,z) = P[\varphi_1 = 1 | U_1 = u,\ V_1 = v,\ A_1],$$

$$(2.13)\qquad P^{(10)}_{n_1,n_2}(u,v,z) = P[\varphi_1 = \tfrac{1}{2} | U_1 = u,\ V_1 = v,\ A_1].$$

Now it can be seen that

$$(2.14)\qquad P^{(11)}_{n_1,n_2}(u,v,z) = C_1(n_1,n_2)/B(n_1,n_2),$$

$$(2.15)\qquad P^{(10)}_{n_1,n_2}(u,v,z) = C_0(n_1,n_2)/B(n_1,n_2),$$

where

(2.16)

$$C_1(n_1,n_2) = n_1(n_1-1)[1-\{F_1(v) - F_1(u)\}]^{n_1-2}[1-\{F_2(v) - F_2(u)\}]^{n_2} f_1(u)f_1(v),$$

(2.17)

$$C_2(n_1,n_2) = n_2(n_2-1)[1-\{F_1(v) - F_1(u)\}]^{n_1}[1-\{F_2(v) - F_2(u)\}]^{n_2-2} f_2(u)f_2(v),$$

(2.18)

$$C_0(n_1,n_2) = n_1 n_2[1-\{F_1(v) - F_1(u)\}]^{n_1-1}[1-\{F_2(v) - F_2(u)\}]^{n_2-1}$$

$$\{f_1(u)f_2(v) + f_2(u)f_1(v)\},$$

$$(2.19)\qquad B(n_1,n_2) = C_1(n_1,n_2) + C_2(n_1,n_2) + C_0(n_1,n_2).$$

Let

$$p_i = \lim_{n\to\infty} n_i/(n_1 + n_2) \quad , \quad i = 1,2. \tag{2.20}$$

We assume that $0 < p_1 < 1$.

Theorem 2.1. Suppose z is a point of continuity of both f_1 and f_2, and $f_1(z) \cdot f_2(z) > 0$. Then for almost all z (under f_1 or f_2)

$$\pi^{(1)}(z) \equiv \lim_{n\to\infty} \pi^{(1)}(z;n_1,n_2)$$

$$= \eta_1 + \tfrac{1}{2}\eta_0, \tag{2.21}$$

where

$$\eta_1 = p_1^2 f_1^2(z)/\{p_1 f_1(z) + p_2 f_2(z)\}^2, \tag{2.22}$$

$$\eta_0 = 2p_1 p_2 f_1(z) f_2(z)/\{p_1 f_1(z) + p_2 f_2(z)\}^2 \tag{2.23}$$

Proof. When u, v $\to$ z and n $\to \infty$.

$$P^{(11)}_{n_1,n_2}(u,v,z) \to \eta_1, \tag{2.24}$$

$$P^{(10)}_{n_1,n_2}(u,v,z) \to \eta_0. \tag{2.25}$$

The desired result now follows from (2.11), (2.10), (2.9), Lemma 2.2, and the dominated convergence theorem.

The limiting PMC's of the one-stage RNN rule are given as follows.

$$\alpha_1^{(1)} \equiv \lim_{n\to\infty} P(\text{Decide } Z \in \pi_2 | Z \in \pi_1)$$

$$= \int [1-\pi^{(1)}(z)] f_1(z)dz \tag{2.26}$$

$$= \int [p_2 f_2(z) f_1(z)/\{p_1 f_1(z) + p_2 f_2(z)\}]dz$$

$$\alpha_2^{(1)} = \lim_{n\to\infty} P(\text{Decide } Z \in \pi_1 | Z \in \pi_2)$$

$$= \int \pi^{(1)}(z) f_2(z)dz.$$

$$= \int [p_1 f_1(z) f_2(z)/\{p_1 f_1(z) + p_2 f_2(z)\}]dz. \tag{2.27}$$

When the training sample is an identified sample from the mixture of π_1 and π_2 with the mixture proportion ξ_1 and ξ_2, respectively, we may take $p_i = \xi_i$ $(i = 1,2)$. Then the limiting value of the total PMC (or, the Bayes' risk) of the one-stage RNN rule is

$$(2.28)\quad r^{(1)} = \int [2\xi_1\xi_2 f_1(z)f_2(z)/\{\xi_1 f_1(z) + \xi_2 f_2(z)\}]dz.$$

If ξ_i's and f_i's were known, the minimum value of the total PMC (or, the risk of a Bayes' rule) is given by

$$(2.29)\quad r^* = \int \min\,[\xi_1 f_1(z),\ \xi_2 f_2(z)]dz$$

It can be seen easily that

$$r^* < r^{(1)} \leq 2r^*$$

See [2]. It may be noted that the result of Theorem 2.1 holds a.e. (μ) for z such that $f_1(z) + f_2(z) > 0$ instead of $f_1(z)\ f_2(z) > 0$.

3. Limiting PMC's of the M-stage RNN rule.

Let $\pi^{(M)}(z;n_1,n_2)$ be the conditional probability that the M-stage RNN rule classifies Z into π_1 given $Z = z$. Let

$$(3.1)\quad \pi^{(M)}(z) = \lim_{n\to\infty} \pi^{(M)}(z;n_1,n_2)$$

Recall the definition of φ_i given in (2.8). Then

$$\begin{aligned} \pi^{(M)}(z;n_1,n_2) &= P[\varphi_1 = 1|Z=z] \\ &+\sum_{i=2}^{M} P[\varphi_1 = \tfrac{1}{2},\dots,\varphi_{i-1}= \tfrac{1}{2},\varphi_i = 1|Z=z] \\ (3.2)\qquad &+ \tfrac{1}{2}\cdot P[\varphi_1 = \tfrac{1}{2},\dots,\varphi_M = \tfrac{1}{2}|Z=z]. \end{aligned}$$

Now

$$P[\varphi_1 = \tfrac{1}{2},\dots,\varphi_{i-1} = \tfrac{1}{2},\varphi_i = \tfrac{1}{2}|Z=z]$$

$$= \prod_{j=2}^{i} P[\varphi_j = \tfrac{1}{2}|\varphi_1 = \tfrac{1}{2},\dots,\varphi_{j-1} = \tfrac{1}{2},Z=z]P[\varphi_1 = \tfrac{1}{2}|Z=z]$$

$$(3.3) \quad P[\varphi_1 = \tfrac{1}{2},\ldots,\varphi_{i-1} = \tfrac{1}{2},\varphi_i = 1 | Z=z]$$
$$= P[\varphi_i = 1 | \varphi_1 = \tfrac{1}{2},\ldots,\varphi_{i-1} = \tfrac{1}{2}, Z=z]$$
$$\prod_{j=2}^{i-1} P[\varphi_j = \tfrac{1}{2} | \varphi_1 = \tfrac{1}{2},\ldots,\varphi_{j-1} = \tfrac{1}{2}, Z=z].$$

$$(3.4) \quad P(\varphi_i = \tfrac{1}{2} | Z=z).$$

We shall show that under certain conditions

$$\lim_{n\to\infty} P[\varphi_i = \tfrac{1}{2} | \varphi_1 = \tfrac{1}{2},\ldots,\varphi_{i-1} = \tfrac{1}{2}, Z=z]$$
$$(3.5) \quad = \lim_{n\to\infty} [\varphi_i = \tfrac{1}{2} | Z=z] = \eta_0.$$

and

$$\lim_{n\to\infty} P[\varphi_i = 1 | \varphi_1 = \tfrac{1}{2},\ldots,\varphi_{i-1} = \tfrac{1}{2}, Z=z]$$
$$(3.6) \quad = \lim_{n\to\infty} P[\varphi_1 = 1 | Z=z] = \eta_1,$$

where η_0 and η_1 are given by (2.22) and (2.23). Then

$$(3.7 \quad \pi^{(M)}(z) = \eta_1 \sum_{i=0}^{M-1} \eta_0^i + \tfrac{1}{2}\eta_0^M$$

Suppose $\varphi_1 = \frac{1}{2}$. Delete the observations corresponding to U_1 and V_1 from the pooled training sample. Denote the remaining n_1-1 X-observations and n_2-1 Y-observations by $X_i^{(2)}(i=1,\ldots,n_1-1)$ and $Y_j^{(2)}(j=1,\ldots,n_2-1)$, respectively, maintaining the orders of the original subscripts.

Lemma 3.1. Given $Z=z$, $\varphi_1 = \frac{1}{2}$, $U_1 = u_1$, $V_1 = v_1$, the conditional distribution of $X_i^{(i)}$'s and $Y_j^{(2)}$'s is given as follows.

(i) $X_i^{(2)}$'s and $Y_j^{(2)}$'s are mutually independent.

(ii) The density of $X_i^{(2)}$ is

$$(3.8) \quad f_1^{(2)}(x) = f_1(x) / [1-\{F_1(v_1) - F_1(u_1)\}],$$

on the complement of $[u_1,v_1]$.

(iii) The density of $Y_j^{(2)}$ is

$$(3.9) \quad f_2^{(2)}(y) = f_2(y)/[1-\{F_2(v_1 - F_2(v_1)\}]$$

on the complement of $[u_1,v_1]$.

Lemma 3.1 can be extended in similar lines inductively to the following. Suppose $\varphi_j = \frac{1}{2}$ $(j = 1,\dots,i-1)$. Delete the observations corresponding to U_j and V_j $(j = 1,\dots, i-1)$ and denote the remaining $n_1 - i+1$ X observations and $n_2 - i+1$ Y-observations by $X_1^{(i)}(r = 1,\dots, n_1-i+1)$ and $Y_r^{(i)}$ $(r = 1,\dots,n_2-i+1)$, respectively, maintaining the order of the original subscripts.

<u>Lemma 3.2</u>. Given $Z=z$, $U_j = u_j$, $V_j = v_j$, $\varphi_j = \frac{1}{2}$ $(j=1,\dots,i-1)$ the conditional distribution of $X_r^{(i)}$'s and $Y_r^{(i)}$'s is given as follows.

(i) $X_r^{(i)}$'s and $Y_r^{(i)}$'s are mutually independent.

(ii) The density of $X_r^{(i)}$ is

$$(3.10) \qquad f_1^{(i)}(x) = f_1^{(i-1)}(x)/ [1 - \{F_1^{(i-1)}(v_{i-1}) - F_1^{(i-1)}(u_{i-1})\}]$$

on the complement of $[u_{i-1}, v_{i-1}]$, where $F_1^{(i-1)}$ is the c.d.f corresponding to $f_1^{(i-1)}$, defined inductively by (3.10) and (3.8).

(iii) The density of $Y_r^{(i)}$ is

$$(3.11) \qquad f_2^{(i)}(y) = f_2^{(i-1)}(y)/[1 - \{F_2^{(i-1)}(v_{i-1}) - F_2^{(i-1)}(u_{i-1})\}]$$

on the complement of $[u_{i-1}, v_{i-1}]$, where $F_2^{(i-1)}$ is the c.d.f corresponding to $f_2^{(i-1)}$, defined inductively by (3.11) and (3.9).

The above two lemmas can be proved following the line of proof of a similar theorem in one-sample case given by Anderson [1]. Their straightforward but lengthy proofs are omitted.

<u>Theorem 3.1</u>. Under the assumptions of Theorem 2.1 the limiting probability of classifying Z into π_1 using the M-stage RNN rule, given $Z = z$, is given by (3.7), for almost all z (under f_1 or f_2).

Proof. As in Section 2 the conditonal probabilities of $\varphi_i = 1$ and $\varphi_i = \frac{1}{2}$, given $Z = z$, $U_j = u_j$, $V_j = v_j$ $(j = 1,\ldots,i)$ and $\varphi_j = \frac{1}{2}$ $(j = 1,\ldots,i-1)$, are respectively given by $C_1^{(i)}/B^{(i)}$ and $C_0^{(i)}/B^{(i)}$, where $C_1^{(i)}, C_2^{(i)}, C_0^{(i)}, B^{(i)}$ are obtained from C_1, C_2, C_0, B, respectively, (see (2.16) - (2.19)) after replacing n_1, n_2, u, v, f_1, f_2 by n_1-i+1, n_2-i+1, u_i, v_i, $f_1^{(i)}$, $f_2^{(i)}$, respectively. Note that if f_j's are continuous at z and u_i and v_i tend to z, then $f_j^{(i)}(u_i) \to f_j(z)$, $f_j^{(i)}(v_i) \to f_j(z)$ $(j = 1,2)$ as $n \to \infty$. Then the limiting values of $C_1^{(i)}/B^{(i)}$ and $C_0^{(i)}/B^{(i)}$ are respectively η_1 and η_0. Now (3.5) and (3.6) follow from Lemma 2.2 and the dominated convergence theorem. As in Theorem 2.1 we can introduce the sets A_i (see after (2.8)) and argue as in (2.9) - (2.11). Now (3.7) follows from (3.2) - (3.6).

The limiting PMC's of the M-stage RNN rule are given as follows.

$$\alpha_1^{(M)} = \lim_{n\to\infty} [\text{M-stage RNN rule decides } Z \in \pi_2 | Z \in \pi_1]$$

$$= \int [1 - \pi^{(M)}(z)]\, f_1(z)dz, \tag{3.12}$$

$$\alpha_2^{(M)} = \lim_{n\to\infty} [\text{M-stage RNN rule decides } Z \in \pi_1 | Z \in \pi_2]$$

$$= \int \pi^{(M)}(z) f_2(z)dz. \tag{3.13}$$

Again in the case of a training sample from the mixed population we may take $p_i = \xi_i (i = 1,2)$. Then the limiting value of the total PMC (or, the Bayes' risk) for the M-stage RNN rule is

$$r^{(M)} = \xi_1\alpha_1^{(M)} + \xi_2\alpha_2^{(M)}$$

$$= \int [\{(\xi_1\xi_2 f_1(z)f_2(z)/(\xi_1 f_1(z) + \xi_2 f_2(z))\} \sum_{i=0}^{M-1} \eta_0^i$$

$$+ \tfrac{1}{2}\eta_0^M (\xi_1 f_1 + \xi_2 f_2)]dz.$$

Now

$$r^{(M)} - r^{(M-1)}$$

$$= -\tfrac{1}{2}\int [\eta_0^{M-1}(\xi_1 f_1(z) - \xi_2 f_2(z))^2/(\xi_1 f_1(z) + \xi_2 f_2(z))dz$$

(3.15)

$$\le 0.$$

Moreover,

$$r^{(\infty)} \equiv \lim_{M\to\infty} r^{(M)}$$

$$= \int [\xi_1\xi_2 f_1(z)f_2(z) \cdot \{\xi_1 f_1(z) + \xi_2 f_2(z)\}]$$

(3.16)

$$[\xi_1^2 f_1^2(z) + \xi_2^2 f_2^2(z)]^{-1}dz$$

It can be seen that

(3.17) $\quad r^* < r^{(\infty)},$

where r^* is the minimum Bayes' risk as given by (2.29).

4. Estimation of PMC's of the one one-stage RNN rule.

We shall estimate the PMC's of the one-stage RNN rule by the deleted counting method described as follows. Let

(3.18) $\quad \Psi_x^{(i)} = 1 - \varphi_1(X_i; X_j\text{'s}, Y_\ell\text{'s};\ j \neq i),$

(3.19) $\quad \Psi_y^{(k)} = \varphi_1(Y_k;\ X_j\text{'s},\ Y_\ell\text{'s};\ \ell \neq k),$

where φ_1 is given by (2.8). Let

(3.20) $\quad p_x(n_1,n_2) = \sum_{i=1}^{n_1} \Psi_x^{(i)}/\ n_1,$

(3.21) $\quad p_y(n_1,n_2) = \sum_{k=1}^{n_2} \Psi_y^{(k)}/n_2\ ,$

(3.22) $\quad p(n_1,n_2) = [n_1 p_x(n_1,n_2) + n_2 p_y(n_1,n_2)]/\ (n_1 + n_2)$

Then p_x and p_y can be used as estimates of the PMC's. Note that

(3.23) $\quad Ep_x(n_1,n_2) = \int[1 - \pi^{(1)}(z;n_1-1,n_2)]f_1(z)dz,$

(3.24) $\quad Ep_y(n_1,n_2) = \int \pi^{(1)}(z;n_1,n_2-1)f_2(z)dz.$

Order the observations in the training sample and denote the number of X-runs and the number of Y-runs by r_1 and r_2, respectively. Then it can be seen that

(3.25) $\quad n_1p_x(n_1,n_2) = r_1 + \delta_1,$

(3.26) $\quad n_2p_y(n_1,n_2) = r_2 + \delta_2,$

where $|\delta_i| \leq 1$ $(i = 1,2)$; δ_i's are the contributions arising from the extreme observations. Let r be the total number of runs. Thus, using (2.26) and (2.27) , we get

(3.27) $\quad \lim_{n\to\infty} E(r_1/n_1) = \alpha_1^{(1)},$

(3.28) $\quad \lim_{n\to\infty} E(r_2/n_2) = \alpha_2^{(1)},$

and

$$\lim_{n\to\infty} E(r/(n_1 + n_2)) = p_1\alpha_1^{(1)} + p_2\alpha_2^{(1)}$$

(3.29) $$= \int [2p_1p_2f_1(z)f_2(z)/ \{p_1f_1(z) + p_2f_2(z)\}]dz$$

The result (3.29) is well-known in the theory of runs and it was derived by Wald and Wolfowitz [5]. Now the result (3.29) may be used to give short proofs of (2.26) and (2.27) after noting the fact that $|r_1 - r_2| = 0$ or 1. Similar estimates of the PMC's of the multistage RNN rules can be obtained; however, they can't be reduced easily as in (3.25) and (?.26).

Note 1. Suppose the c.d.f of Z is F. For the one-stage RNN rule the conditional probability of classifying Z into π_2 given the training sample is derived as follows. Let

(3.30) $\quad T_1 < T_2 < \ldots < T_{n_1 + n_2}$

be the ordered values of the observations in the training sample. Write

$$(3.31) \qquad \theta_i = \begin{cases} 0, & \text{if } T_i \text{ is an X-observation} \\ 1, & \text{if } T_i \text{ is an Y-observation} \end{cases}$$

Then the conditional probability of classifying Z into π_2 using the one-stage RNN rule given the training sample is

$$(3.32) \qquad \theta_1 F(T_1) + \tfrac{1}{2}\sum_{i=1}^{n_1+n_2} [F(T_{i+1}) - F(T_i)]\,(\theta_{i+1} + \theta_i)$$
$$+ [1 - F(T_{n_1+n_2})]\theta_{n_1+n_2}$$

The behavior of (3.32) is under investigation.

Note 2. It will be quite useful to compare the NN rule and the different RNN rules when n is small and under specific F_1 and F_2. Monte Carlo studies on these problems will be reported later.

Note 3. The results in Section 2 are taken from the Ph.D. thesis of the second author [4] and modified suitably.

REFERENCES

[1] Anderson, T. W. (1966). Some nonparametric multivariate procedures based on statistically equivalent blocks. Proc. 1st Internat. Symp. Multivariate Anal. Ed. P. R. Krishnaiah. Academic Press, New York.

[2] Cover T. M. and Hart P. E. (1967). Nearest neighbor pattern classification. IEEE Trans. Inform. Theory IT - 13, 21-26.

[3] Fix, E. and Hodges, J. L. (1951). Nonparametric discrimination: Consistency properties. U.S. Air Force School of Aviation Medicine, Report No. 4. Randolph Field, Texas.

[4] Lin, H. E. (1976). One some aspects of the classification problem. Tech. Report 276, School of Statistics, University of Minnesota, Minneapolis.

[5] Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. Ann. Math. Statist., 11, 147-162.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER UMN/DTS/TR-285	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Nearest Neighbor Rules for Statistical Classification Based on Ranks		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Somesh Das Gupta and Hsien Elsa Lin		8. CONTRACT OR GRANT NUMBER(s) DAAG-29-76-0038
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Theoretical Statistics University of Minnesota Minneapolis, MN 55455		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE May 5, 1977
		13. NUMBER OF PAGES 14
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE NA

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Classification, rank nearest neighbor rule, multi-stage rule, asymptotic probability of misclassification, estimation of error probabilities.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Some rank-based rules for classifying an observation into one of two populations, from which training samples are available, are suggested. The limiting values of their probabilities of misclassification are derived as the sizes of the training samples tend to infinity. The nearest neighbor idea is used in terms of ranks and two-sided neighbors to define the rank nearest neighbor rule. A multi-stage version is also suggested (in case of ties) and studied.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)